DOCUMENT RESUME

ED 308 229 TM 013 608

AUTHOR	Wise, Lauress L.; And Others
TITLE	Item Position Effects for Test of Word Knowledge and Arithmetic Reasoning.
INSTITUTION	American Institutes for Research, Washington, DC.
SPONS AGENCY	Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
PUB DATE	Mar 89
CONTRACT	MDA903-86-C-0373
NOTE	37p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Armed Forces; *Computer Assisted Testing; Item Analysis; Latent Trait Theory; *Mathematics Skills; *Military Personnel; Recruitment; Statistical Analysis; Testing Programs; *Test Items; *Vocabulary Skills
IDENTIFIERS	Arithmetic Reasoning; Army; *Computerized Adaptive Screening Test; Item Parameters; *Item Position (Tests); Word Knowledge

ABSTRACT

The effects of item position on item statistics were studied in a large set of data from tests of word knowledge (WK) and arithmetic reasoning (AR). Position effects on item response theory (IRT) parameter estimates and classical item statistics were also investigated. Data were collected as part of a project to refine the Army's Computerized Adaptive Screening Test (CAST). The CAST is an adaptively administered battery consisting of a WK subtest and an arithmetic reasoning (AR) subtest. As part of this effort to refine the CAST, 275 new and existing items from the WK and AR subtests were administered to 20,071 Army recruits from five different Army posts. A total of 270 of the items to be calibrated for each subtest was divided into six non-overlapping sets of 45 items each. The remaining five items were included in all six forms as potential anchors should subsequent equating prove necessary. Item statistics were computed separately for forward and reversed versions of each form. IRT parameters and classical parameters were determined. Estimates for both parameters varied significantly with item position. The variation was not generally related to the characteristics of the item, but was related to the ability of the examinees. There were no significant position effects when average percent passing scores were 75% or higher; position effects were quite pronounced when passing scores were 50% or lower. The primary conclusion drawn is the need to avoid unquestioning adoption of IRT methodology. Including the reversed version of each form prevented the introduction of systematic errors in the IRT parameter estimates. Seven data tables and 11 graphs illustrate the study findings. (SLD)

ED308229

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Lauress L. Wise

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Item Position Effects for Test of Word Knowledge and Arithmetic Reasoning

Lauress L. Wise
Wei Jing Chia
American Institutes for Research

Randolph K. Park
American Association of Medical Colleges

Paper presented at the Annual Meeting of the American Educational Research Association, 28 March 1989, San Francisco, California.

Note: The research reported here was supported by the US Army Research Institute under Contract No. MDA903-86-C-0373. The views and findings expressed in this report are those of the authors and should not be construed as an official Department of Army position or policy.

American Institutes for Research
Washington, DC • Palo Alto, CA • Bedford, MA • Watertown, MA

TM013608

BEST COPY AVAILABLE

Item Position Effects for Test of Word Knowledge and Arithmetic Reasoning

Item response theory (IRT) models describe the probability that an item will be answered correctly as a function of examinee ability and of characteristics of the item in question (Lord, 1980). The probability that an item is answered correctly is assumed to be independent of the context in which the item is placed. This includes an item's position in the test and the nature of the other items in the test.

The absence of context effects is particularly critical for adaptive testing. In adaptive testing, different examinees receive different sets of items which may appear in a wide variety of orderings and groupings. Examinee scores depend heavily on an accurate calibration of the difficulty and other parameters of each item. Consequently, calibration must be independent of item position and other context effects.

In a prior research effort, Wise (1986) demonstrated that important item statistics such as the proportion passing and the item-total biserial correlation may vary systematically with item position. The research reported here extends those results through analysis of a larger set of items with more systematic variation in item position. In addition, position effects on IRT parameter estimates as well as on classical item statistics were investigated.

Data

The data reported here were collected as part of a project to refine the Army's Computerized Adaptive Screening Test (CAST). CAST is an adaptively administered battery consisting of a Word Knowledge (WK) subtest and an Arithmetic Reasoning (AR) subtest. It is used by recruiters to obtain a quick assessment of a candidate's chances of passing entrance requirements on the Armed Services Aptitude Battery (ASVAB). The CAST takes about 15 minutes to administer, while the ASVAB takes 3 hours and is part of a day-long assessment process. It is thus in the best interests for both applicants and recruiters that applicants who have little chance of qualifying be discouraged from going to a Military Enlistment Processing Station (MEPS) or Mobile Enlistment Testing Station (METS) for a lengthy assessment.

The CAST Refinement Project was initiated in 1986 to expand the CAST item bank, to screen new and existing items for fairness to minorities and women, to modify the item selection and scoring procedures to improve accuracy at key points, and to improve the format used for reporting

scores to the applicants and recruiters. As part of this effort, 275 new and existing items from each subtest were administered to large samples of new recruits to provide data for estimating IRT parameters for each item. Fixed-item paper-and-pencil forms were used in this effort. An alternative calibration, in which experimental items were seeded within the operational CAST, was used to check for differences in mode of administration and no significant differences were found (Wise, McHenry, Chia, Szenas, & McBride, 1988).

In the primary calibration effort, the items to be calibrated for each subtest were divided into six forms of 50 items each. A total of 270 of the items to be calibrated for each subtest were divided into six nonoverlapping sets of 45 each, balanced as far as possible with respect to expected difficulty and old versus new items. The remaining 5 items were included in all six forms as potential anchors should subsequent equating prove necessary. For each of the six forms, an alternate version was created by reversing the order of the items.

20,071 new recruits from five different Army posts each completed one WK form and one AR form. At each testing site, each of 12 different versions of each subtest was alternately distributed to maximize the comparability of the examinee samples for each form. Table 1 shows the number of examinees completing each form of each subtest.

Insert Table 1 about here.

Method

Item statistics were computed separately for the forward and reversed versions of each form. IRT parameter estimates were computed using the BILOG program (Mislevy & Bock, 1981) using default option priors for the *a* and *c* parameters, "floating" priors for the *b* parameters, and excluding "not reached" items from the analysis. In addition, the proportion passing, the proportion omitting or not reaching each item, and the Clemans-Brodgen variant of the item-total biserial correlation (described in Lord & Novick, 1968, page 340-341) were also analyzed. In a separate analysis, the LOGIST program (Wood, Wingersky, & Lord, 1976) was also used to obtain IRT parameter estimates and found to give virtually identical results (Wise, et al., 1988).

In estimating the IRT parameters for each version of each form, the item and examinee parameters were scaled so that the examinee ability distribution would have a mean of zero and unit variance. Comparisons

with ASVAB scores and analyses of other characteristics for the different samples receiving each version of each form revealed no significant differences among these samples. Consequently, further adjustments to these scales were judged unnecessary. (See Wise, et al., 1988 for a more detailed discussion of alternative equating analyses.)

The question of interest in the present analyses was the extent to which the two versions of each form yielded equivalent item parameter estimates. To address this question, we computed the difference between the estimates derived from each version. In these analyses, results were computed for each item in each form and then pooled across the six different forms. Data for the five items common to all six forms were also analyzed by form, leading to an effective pool of 300 items (50 items for each of six forms) on which version difference statistics were computed.

We examined the extent to which version differences in the item statistics were correlated with item position and with other characteristics of the items such as mean difficulty, discrimination, or guessing levels. In subsequent analyses, we divided the examinees completing each form into six ability strata (using number correct scores) and examined the relationship of examinee ability stratum to differences in item parameter estimates generated from the data for subjects in each stratum.

Results

Table 2 shows mean item statistics across the six forward and six reverse versions of the different forms. No significant differences in means or standard deviations were found. Table 3 shows the correlation of differences in each of the item parameter estimates with differences in item position for each of the two subtests. Here, very significant relationships were found. Figures 1 through 8 show plots of the differences in IRT difficulty and slope estimates and in proportion passing and proportion omitting by item position for each of the two subtests. These plots show item position in the reversed form, version 2. The difference in item position is systematically related to version 2 position since forward item 1 was always reversed item 50, forward item 2 was reversed item 49, etc. For the WK items, the relationship of differences in proportion omitting to item position differences showed significant curvilinearity. In all other cases, the relationships were systematic (not resulting from outliers) and linear.

Insert Tables 2 and 3 about here.

The strong relationship of proportion omitting to item position suggested that these tests may have been more speeded than was originally supposed. We examined the percent omitting by item position for each subtest as shown in Figures 9 and 10. The AR forms showed some significant effects of speed, with only about 85 percent of the examinees completing the entire test. The WK forms showed only minimal effects of speed with 96 to 98 percent of the examinees completing the entire test. In any case, the exclusion of "not reached" items in developing IRT estimates should have removed most of the effects of speededness.

Insert Figures 1 through 10 about here.

The plots of item statistic differences by item position (differences) showed greater variation in the difference statistics for larger changes in item position (i.e., items at the beginning or the end of each form) in comparison to differences where the item position change was minimal (i.e., items in the middle of both versions). Further analyses were conducted to see if the degree of change was related to other item characteristics. We computed residual scores for differences in key item statistics removing the effects of item position differences and then computed the correlation of the residualized scores with various item characteristics. Table 4 shows the resulting correlations.

Insert Table 4 about here.

The results in Table 4 show no systematic relationships between the residualized scores and item difficulty, discrimination, or guessing levels. There were, however, significant relationships among the different residual scores. Some of these differences appeared systematic. Consistent with common findings regarding the correlations among IRT parameter estimate errors, positive differences in guessing parameter estimates were associated with positive differences in slope and difficulty estimates. In other cases, the relationships were more difficult to interpret. Positive differences in the slope parameters were associated with positive differences in the difficulty parameters for the WK items and with negative differences in the difficulty parameters for the AR items. In general, these results indicate a clear need for multivariate analyses of differences in the IRT parameter estimates, but no clear relationships between item characteristics and the differences in item parameter estimates beyond those predicted by item position differences.

In a final set of analyses, we examined the extent to which item position effects were associated with examinee ability levels. We focused these analyses on the WK forms, because they were significantly less speeded in comparison to the AR items. We divided the examinees into six roughly equal strata on the basis of their number correct scores. Table 5 shows the number of examinees in each stratum for each of the six forms.

Because of the relatively small sample sizes within each of the ability by version cells, we did not attempt to compute IRT estimates for each group. The IRT difficulty estimate and the proportion passing were closely linked in all of the analyses that were conducted, so we concentrated on analysis of the proportion passing statistics. In these analyses, all omitted responses were excluded, so the proportion passing estimates for each item were based only on those attempting the item. This approach was a relatively extreme attempt to eliminate effects associated with any degree of speededness.

In these analyses, we reversed the direction of the item position shifts so that it was always positive (here indicating that the item moved forward) and reversed the sign of the difference statistics accordingly. (In the preceding analyses, positive shifts were positive when the item moved toward the end of the test. This shift in our perspective enabled us to draw the nice graphic that you will see in a moment.) We divided the items in each form into five groups on the basis of the degree of item position difference between the two versions of the form (0-9 positions, 10-19 positions, etc.).

Table 6 shows the average proportion passing for each examinee and position shift level. The different ability groups were relatively evenly spaced with respect to average percent correct scores. These results also show that the items in the 30-39 position shift level had greater average difficulty in comparison to the other items, systematically across all ability levels.

Table 7 shows the average difference in the proportion passing for each combination of examinee ability level and item position difference level. This information is presented graphically in Figure 11. Table 7 also shows the mean of the proportion passing differences and the correlation of these differences with item position difference levels for each ability level. These results indicate that item position was significantly related to differences in percent passing for examinees in the two lower ability groups and not for examinees in the remaining four higher level ability groups. The degree of relationship varied systematically, but not linearly, with examinee ability. The degree of item position effect appears to increase exponentially with decreasing examinee ability levels.

Insert Tables 5 through 7 about here.

Discussion

The results of this investigation indicated that estimates for both IRT and classical parameters varied significantly with item position. The variation in item parameter estimates was not generally related to characteristics of the items, but was closely related to examinee ability levels. Among other things, these results underscore the prudence of balancing item position across alternate forms in calibrating items. We were able to pool data from both versions of each form and were reasonably confident that this approach minimized the influence of item position in our experimental forms.

The results also suggest the need for particular caution with tests that are difficult for the examinees on which they are calibrated. When average percent passing scores were 75 percent or higher, there were no significant position effects. When the percent passing dropped to 50 percent or lower, the position effects became quite pronounced.

The primary conclusion that we draw from this research is the need to avoid an unquestioning adoption of IRT methodology. Had we not included the reverse version of each form, we would have introduced systematic errors in our IRT parameter estimates. A consequence of these errors is that we might significantly overestimate the accuracy of examinee ability estimates and thus place too much confidence in the test score results.

These results do not address the more significant question of the effect of item position or other context effects when items are administered adaptively. In computerized testing where each item is separately timed, item position effects might be quite different. More extensive data from computer administered tests, including response latency information, are required to address this question.

References

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley

Mislevy, R. J., & Bock, R. D. (1981). BILOG--maximum likelihood item analysis and test score: LOGISTIC MODEL. Chicago, IL: International Educational Service.

Wise, L.L. (1986, April). Latent trait models for partially speeded tests. Paper presented at the American Educational Research Association, San Francisco.

Wise L. L., McHenry, J. J., Chia, W. J., Szenas, P. L., & McBride, J. R. (1988). Refinement of the Computer Adaptive Screening Test (CAST) (under review). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (Research Bulletin, 76-6). Princeton, NJ: Educational Testing Service.

Table 1
Number of Examinees Tested on Each Arithmetic Reasoning and Word Knowledge Subtest

	Number Tested	
Form	AR	WK
1F	1790	1777
1R	1635	1617
2F	1786	1783
2R	1622	1612
3F	1693	1717
3R	1628	1571
4F	1734	1762
4R	1647	1645
5F	1650	1650
5R	1649	1675
6F	1631	1685
6R	1605	1577
Total	20070	20071

Table 2
Item Statistics for Forward and Reverse Versions (Based on 300 Items from Subtest, Pooled Across Forms)

Statistic	AR Items			WK Items		
	Forward	Reverse	Difference	Forward	Reverse	Difference
Means						
IRT Slope (A)	1.77	1.71	-.07	1.84	1.82	-.02
IRT Difficulty (B)	.78	.79	.01	-.80	-.84	-.04
IRT Guessing (C)	.16	.15	-.01	.19	.18	-.01
Pct. Correct	.42	.42	.00	.73	.73	.00
Pct. Omit	.05	.05	.00	.01	.01	.00
Biserial	.46	.46	.00	.55	.55	.00
Standard Deviations						
IRT Slope (A)	.48	.51	.59	.62	.61	.42
IRT Difficulty (B)	.99	.99	.71	1.18	1.18	.47
IRT Guessing (C)	.05	.05	.05	.07	.06	.07
Pct. Correct	.18	.18	.13	.20	.20	.03
Pct. Omit	.05	.05	.09	.01	.01	.01
Biserial	.12	.12	.06	.12	.12	.07

Note: The standard deviations shown under the "Difference" column are the standard deviations of the item statistic differences computed across the 300 items (and not the difference in the standard deviations).

Table 3
Correlation of Item Statistic Differences
With Item Position Differences
(Based on 300 Items for Each Subtest)

Statistic	AR Items	WK Items
IRT Slope (A)	.72	.34
IRT Difficulty (B)	.77	.49
IRT Guessing (C)	.23	-.51
Pct. Correct	-.83	-.57
Pct. Omit	.99	.92
Biserial	-.34	.66

Table 4
Correlation of Residualized Item Statistic Differences (Controlling for Item Position Differences) with Other Item Statistics

Statistic	Residual IRT Slope AR	Residual IRT Slope WK	Residual IRT Diffic. AR	Residual IRT Diffic. WK	Residual IRT Guess AR	Residual IRT Guess WK	Residual % Passing AR	Residual % Passing WK
Mean IRT Slope	ns	ns	ns	ns	ns	ns	ns	ns
Mean IRT Diff.	ns	ns	ns	ns	ns	ns	ns	ns
Mean IRT Guess.	ns	ns	ns	ns	ns	ns	ns	ns
Mean % Passing	ns	ns	ns	ns	ns	ns	ns	ns
Mean Biserial	ns	ns	ns	ns	ns	.17**	ns	.19**
Mean % Omit	ns	ns	ns	ns	ns	ns	ns	ns
Omit Difference	ns	ns	ns	ns	ns	ns	ns	ns
Bis. Difference	ns	.37**	ns	.24**	-.24**	-.26**	-.17**	ns
Residual Slope	--	--	-.23**	.54**	.16**	.26**	.15**	-.19**
Residual Diff.	-.23**	.54**	--	--	.47**	.38**	-.81**	-.46**
Residual Guess	.16**	.26**	.47**	.38**	--	--	-.12*	.17**
Residual % Pass.	.15**	-.18	-.81**	-.46**	-.12*	.17**	--	--

Note: * - significant at the .05 level
** - significant at the .01 level

Table 5
Number of Examinees by Form, Version, and Ability Level
(For the WK Forms)

FORMID SCORGRP(RANK FOR VARIABLE TOTSC)

Form/Vers	Low 0	1	2	3	4	High 5	TOTAL
1F	284	314	283	335	279	282	1777
1R	265	238	284	315	229	286	1617
2F	290	341	309	233	339	271	1783
2R	254	255	298	246	326	233*	1612
3F	298	237	331	308	237	306	1717
3R	245	253	291	283	203	296	1571
4F	279	332	292	312	234	313	1762
4R	253	320	268	278	253	273	1645
5F	256	278	297	242	263	314	1650
5R	260	277	344	279	243	272	1675
6F	323	263	289	253	261	296	1685
6R	255	193	271	305	262	291	1577
TOTAL	3262	3301	3557	3389	3129	3433	20071

Table 6
Mean Percent Passing by Change in Item Position for Different Examinee Ability Groups

Change in Item Position	Examinee Ability Group: Low 1	2	3	4	5	High 6
0 - 9	51	68	77	84	89	95
10 -19	45	63	73	81	87	95
20 -29	47	64	73	80	87	95
30 -39	40	54	64	72	81	91
40 -49	48	63	72	79	86	94
Total	46	63	72	79	86	94

Table 7
Mean Change in Percent Passing by Change in Item Position for Different Examinee Ability Groups

Change in Item Position	Examinee Ability Group: Low 1	2	3	4	5	High 6
0 - 9	.000	.003	.004	.001	-.002	.001
10 -19	.025	.009	.008	.006	.002	.002
20 -29	.042	.012	.001	.003	-.002	-.003
30 -39	.052	.024	.006	.005	-.004	-.003
40 -49	.086	.032	.016	.001	.004	.002
Mean	.041	.016	.007	.003	-.000	-.001
r	.45	.21	.06	-.019	.00	.01
(Signif.)	(.0001)	(.0003)	(.30)	(.87)	(.94)	(.87)

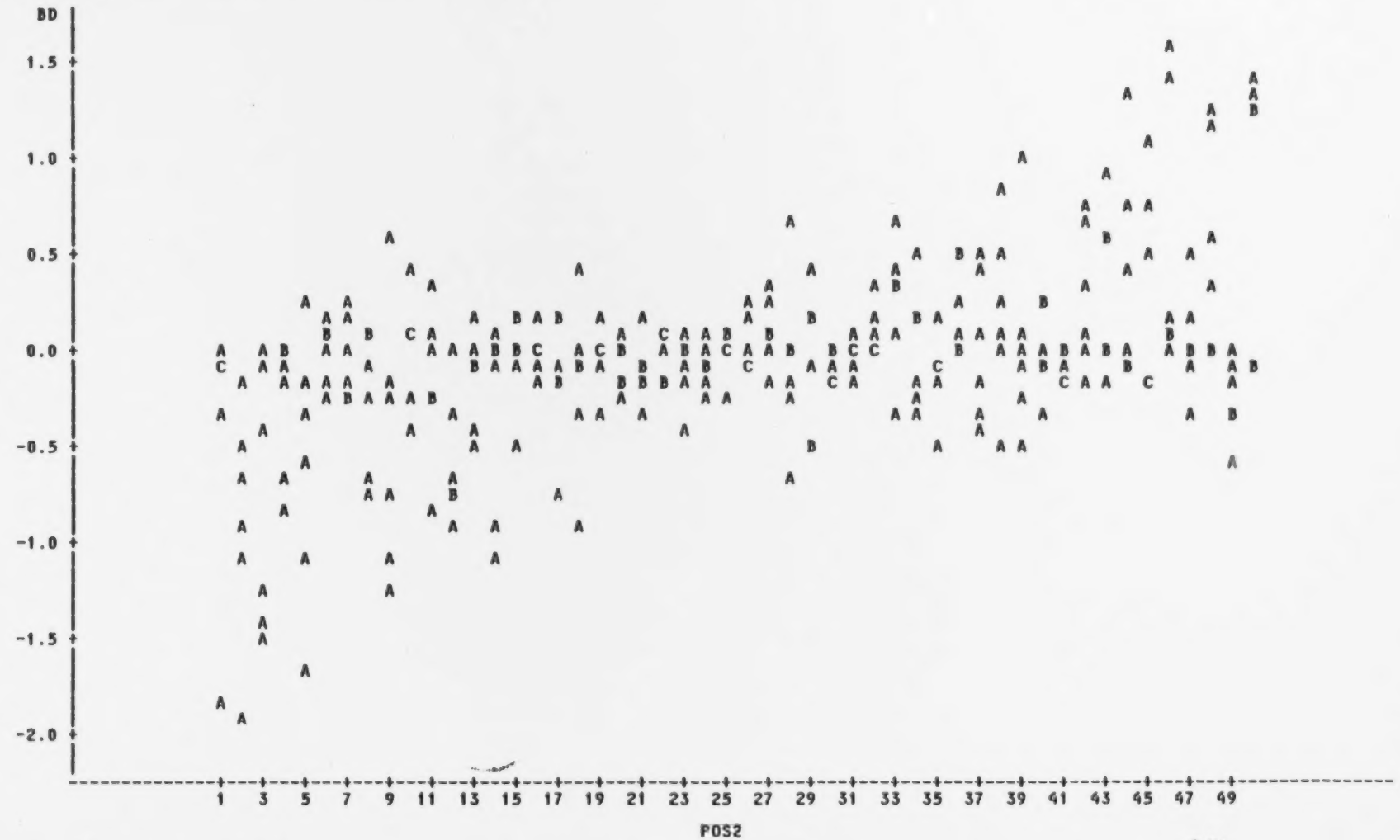

Figure 1. Differences in IRT Difficulty Parameter Estimates (BD) for AR Items by Item Position in Version 2 (POS2).

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

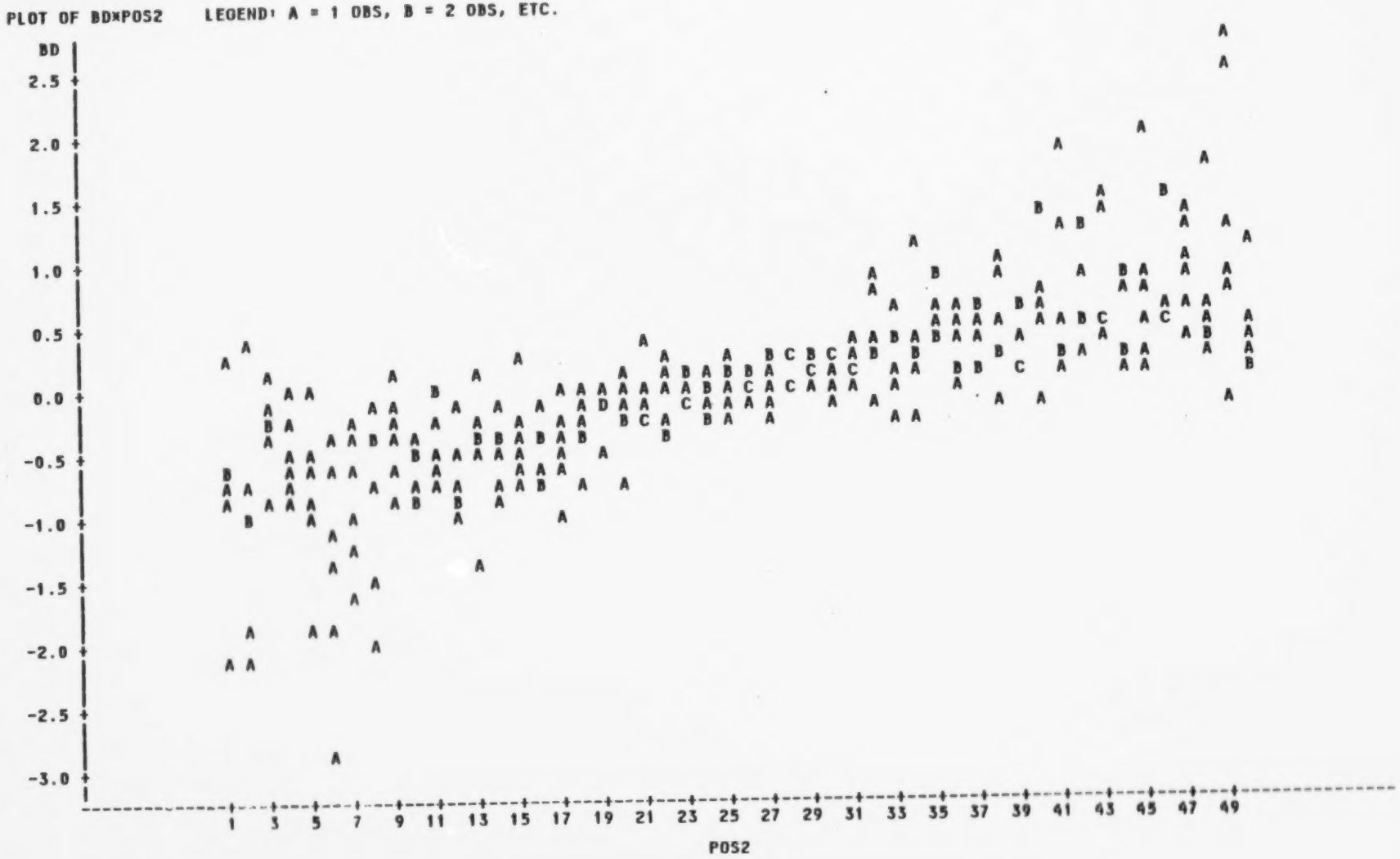

Figure 2. Differences in IRT Difficulty Parameter Estimates (BD) for WK Items by Item Positon in Version 2 (Pos 2).

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

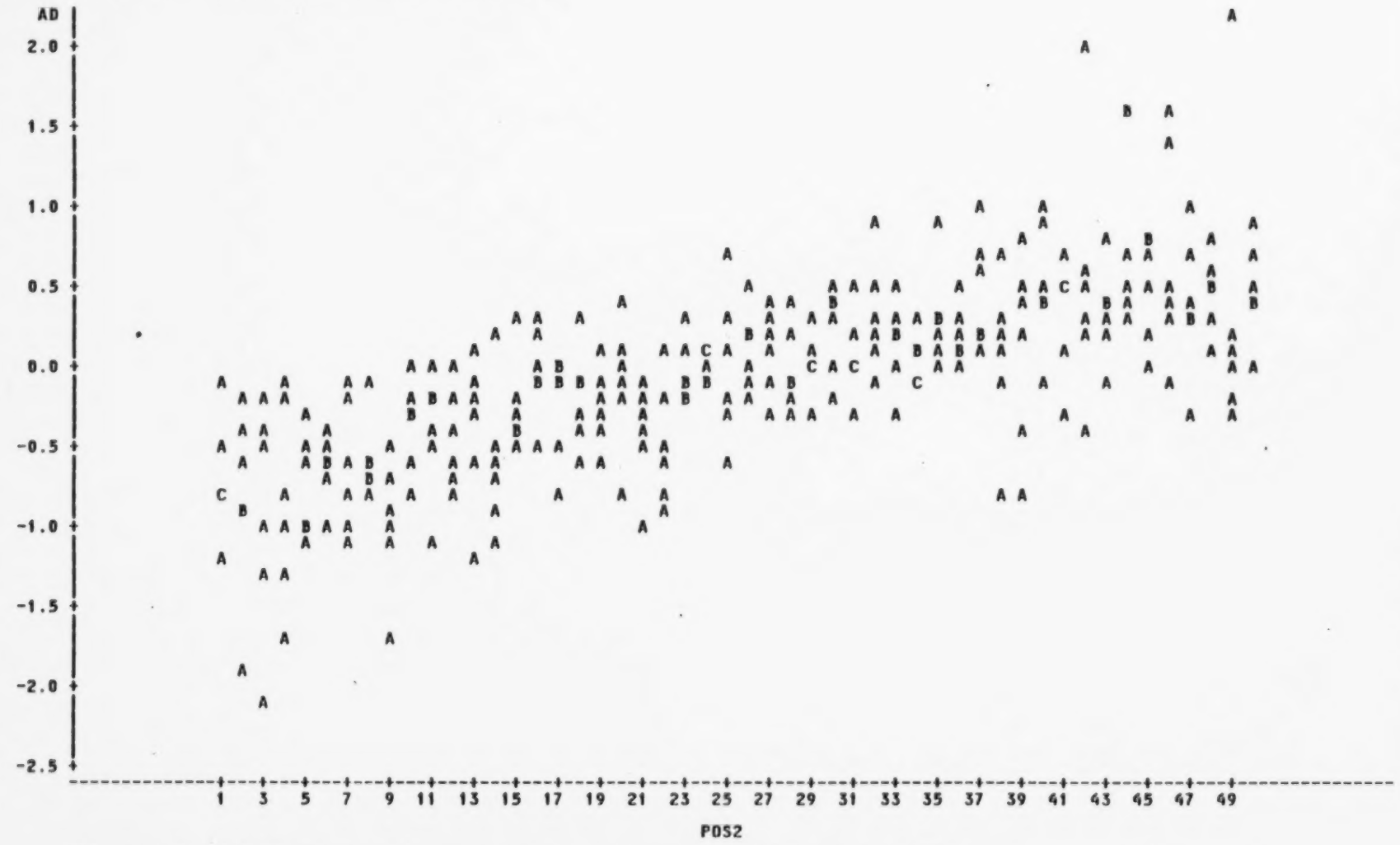

Figure 3. Differences in IRT Slope Parameter Estimates (AD) for AR Items by Item Position in Version 2 (Pos 2).

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

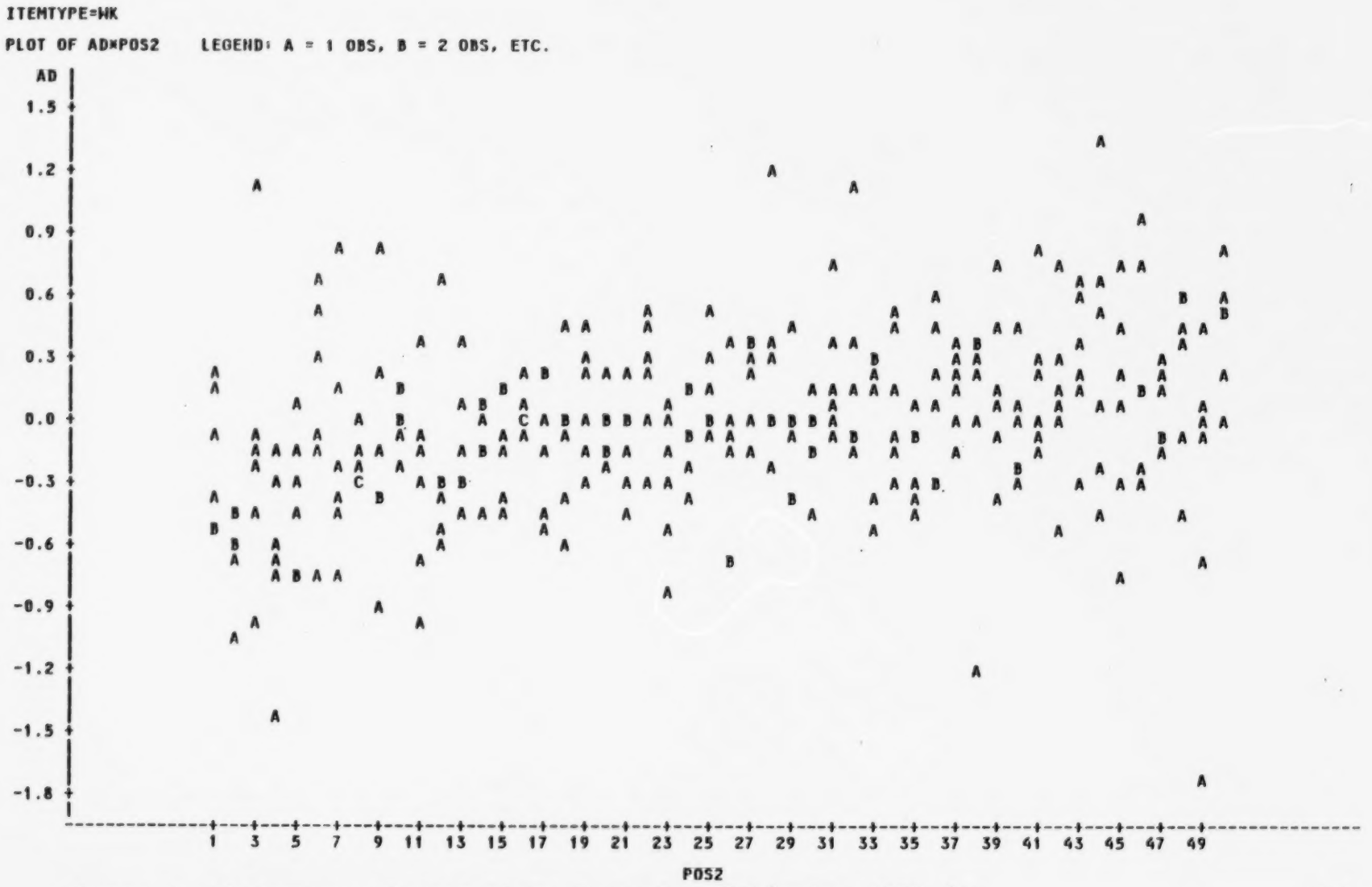

Figure 4. Differences in IRT Slope Parameter Estimates (AD) for WK Items by Item Position in Version 2 (Pos 2).

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

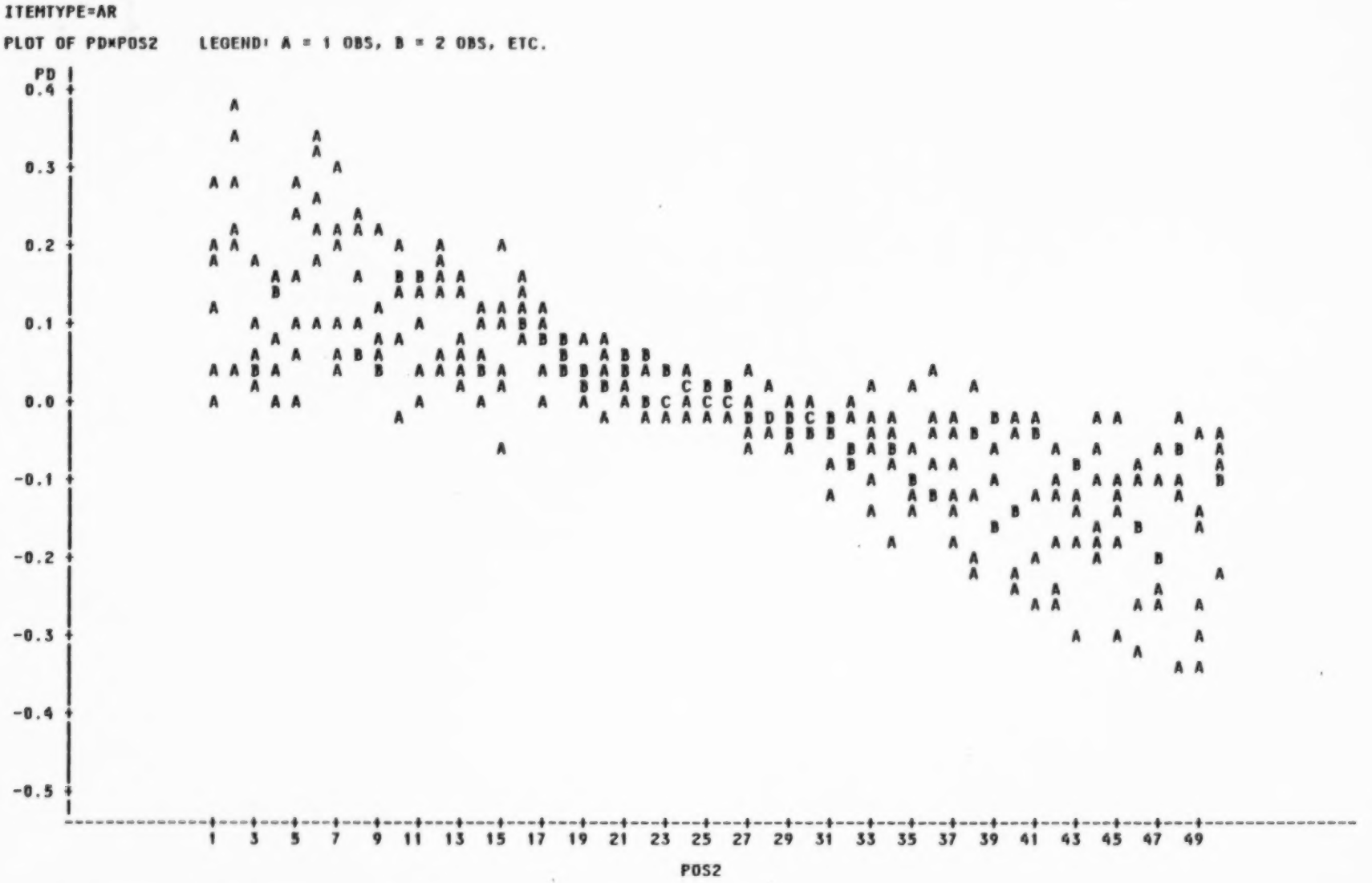

Figure 5. Differences in Percent Passing (PD) for AR Items by Item Position in Version 2 (Pos 2).

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

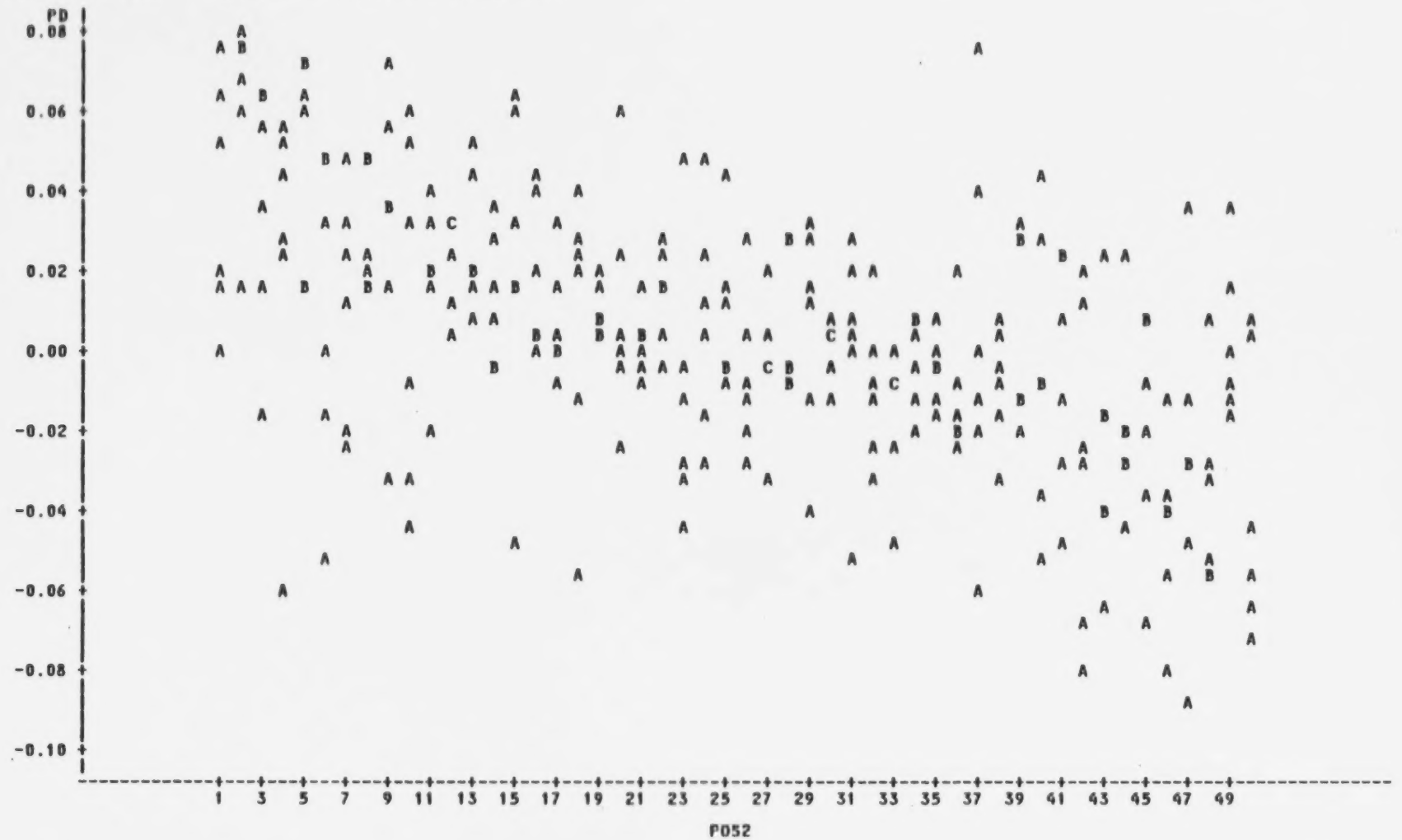

Figure 6. Differences in Percent Passing (PD) for WK Items by Item Position in Version 2 (Pos 2).

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

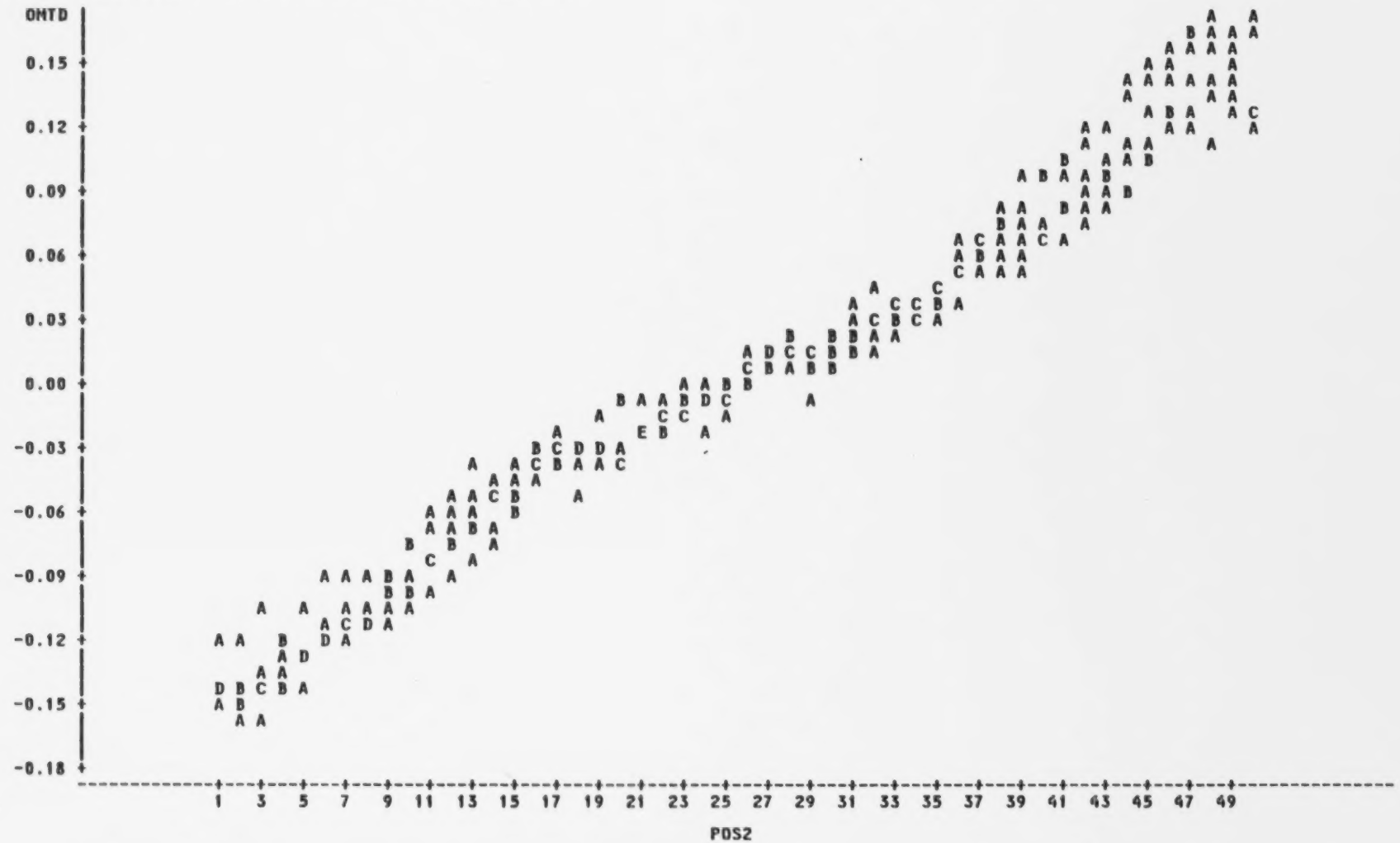

Figure 7. Differences in Percent Omitting (OMTD) for AR Items by Item Position in Version 2 (Pos 2)

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

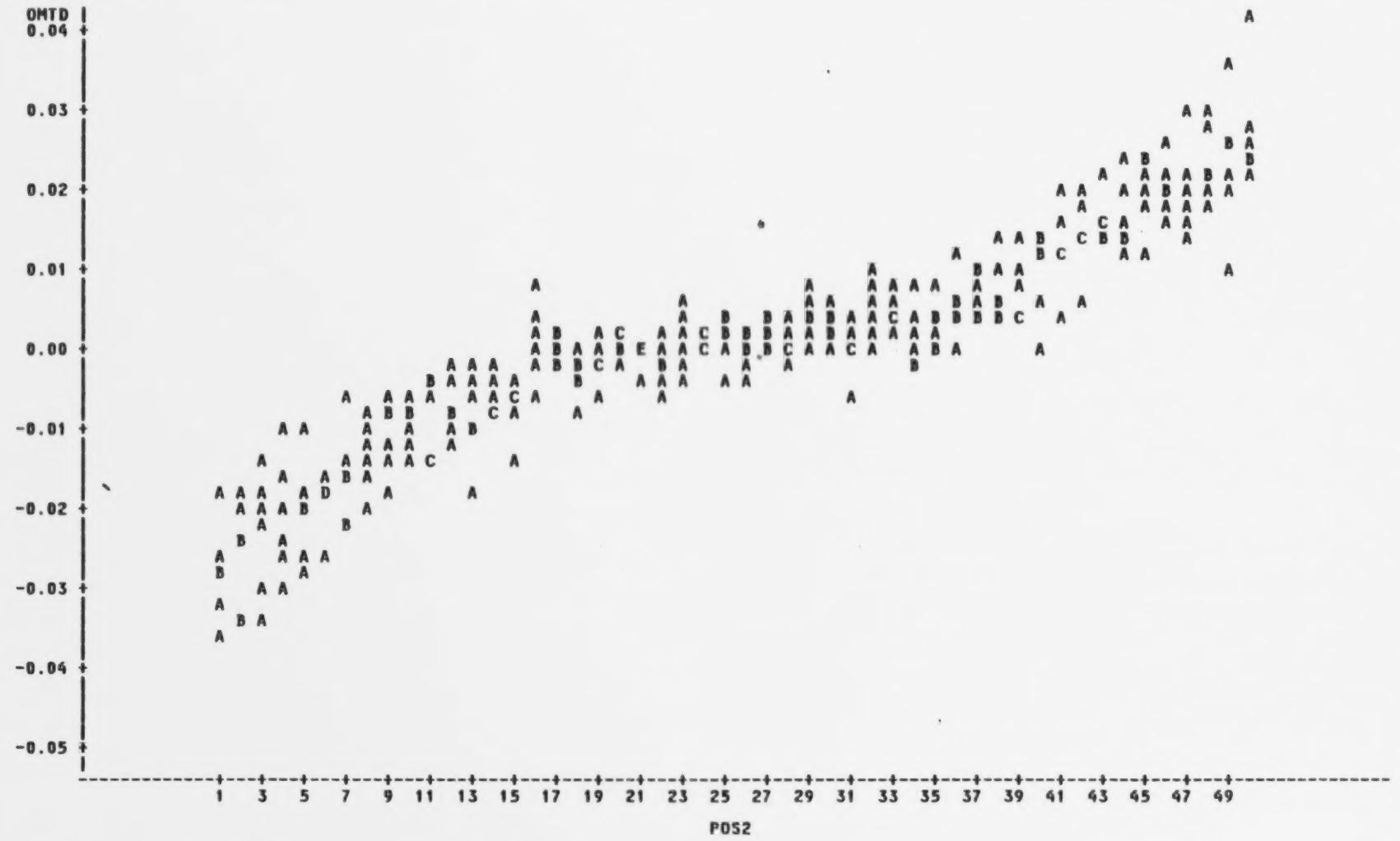

Figure 8. Differences in Percent Omitting (OMTD) for WK Items by Item Position in Version 2 (Pos 2).

Note. Differences were computed as the Version 2 estimate minus the Version 1 estimate.

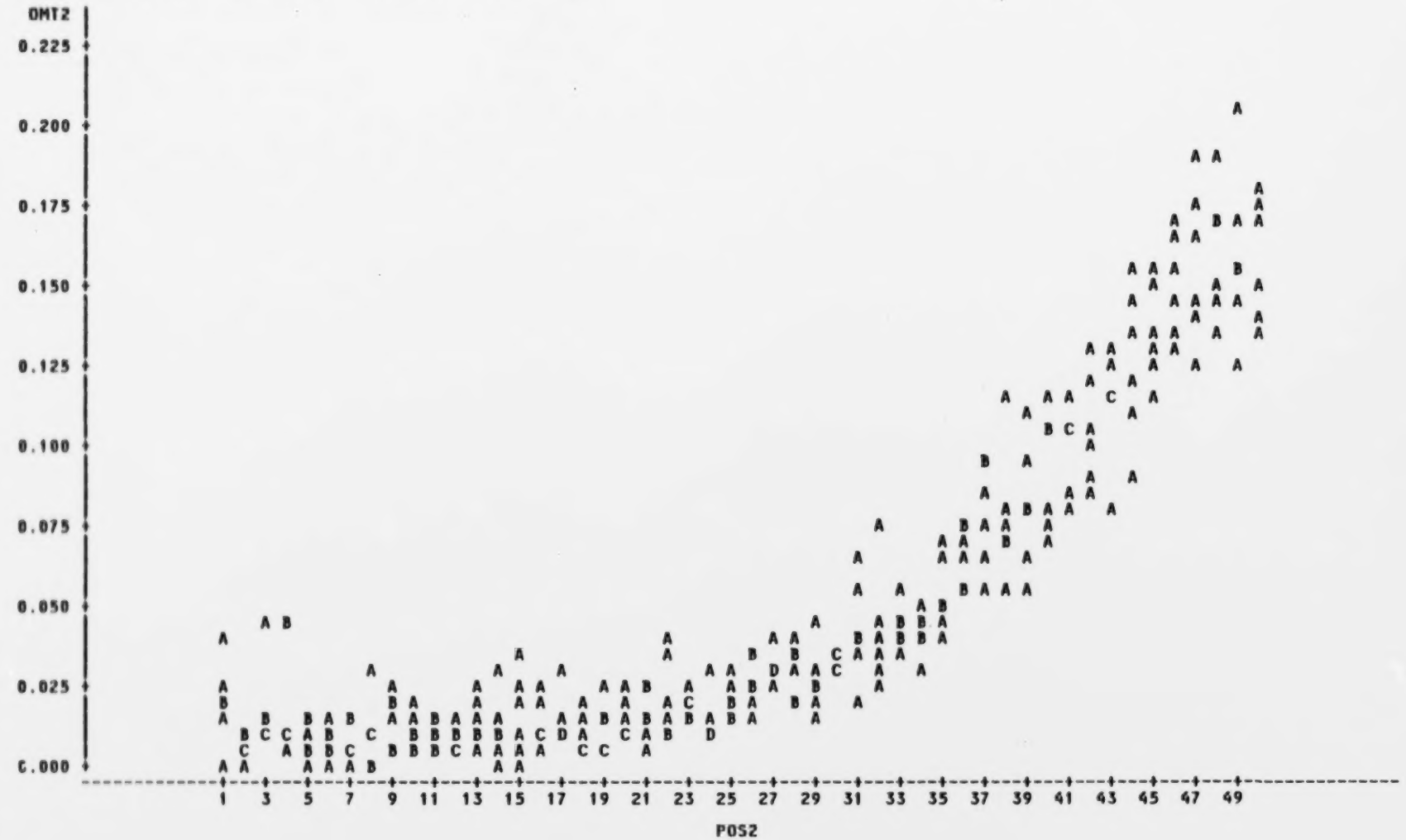

Figure 9. Percent Omitting for AR Items in Version 2 (OMT2) by Item Position.

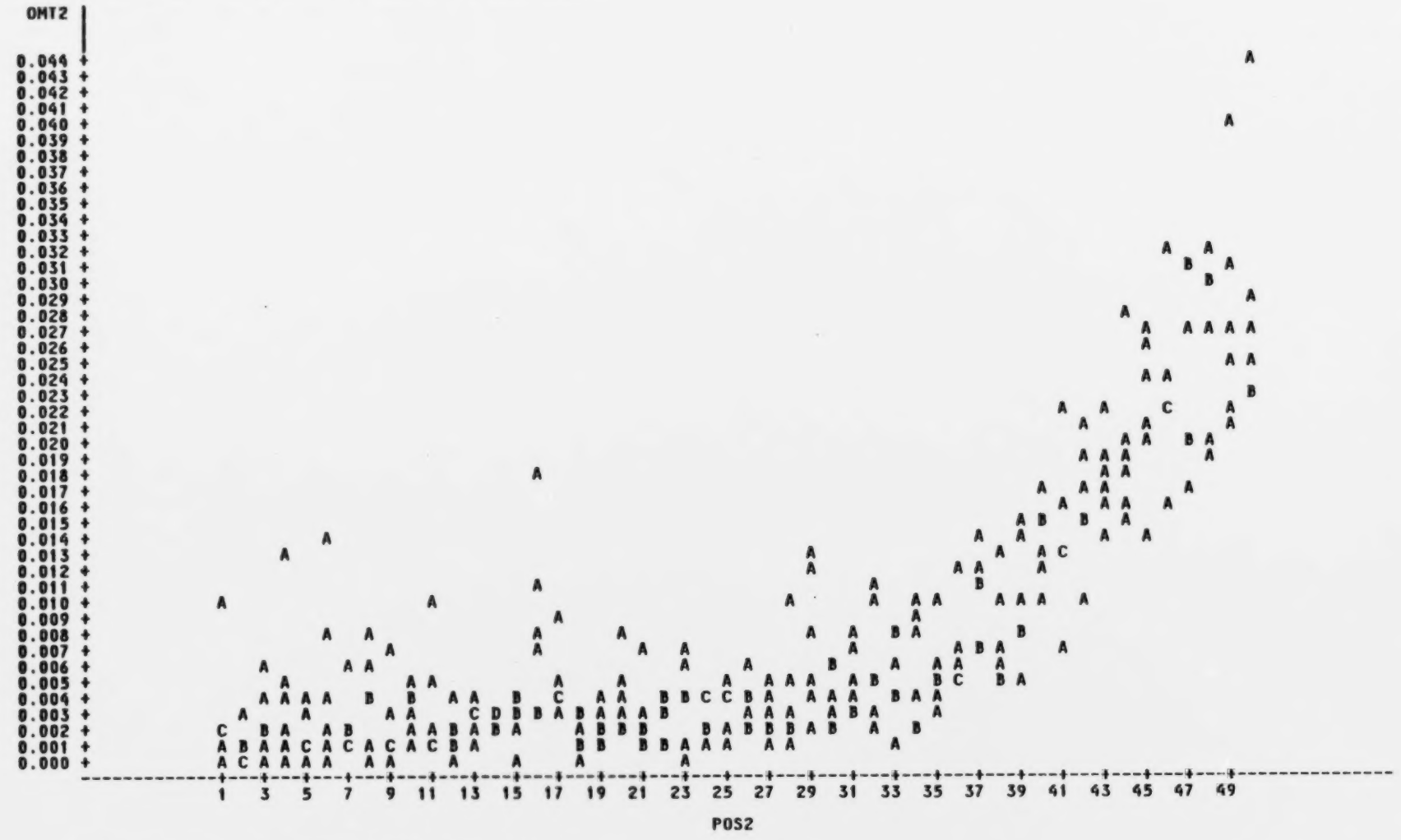

Figure 10. Percent Omitting for Wk Items in Version 2 (OMT2) by Item Position.

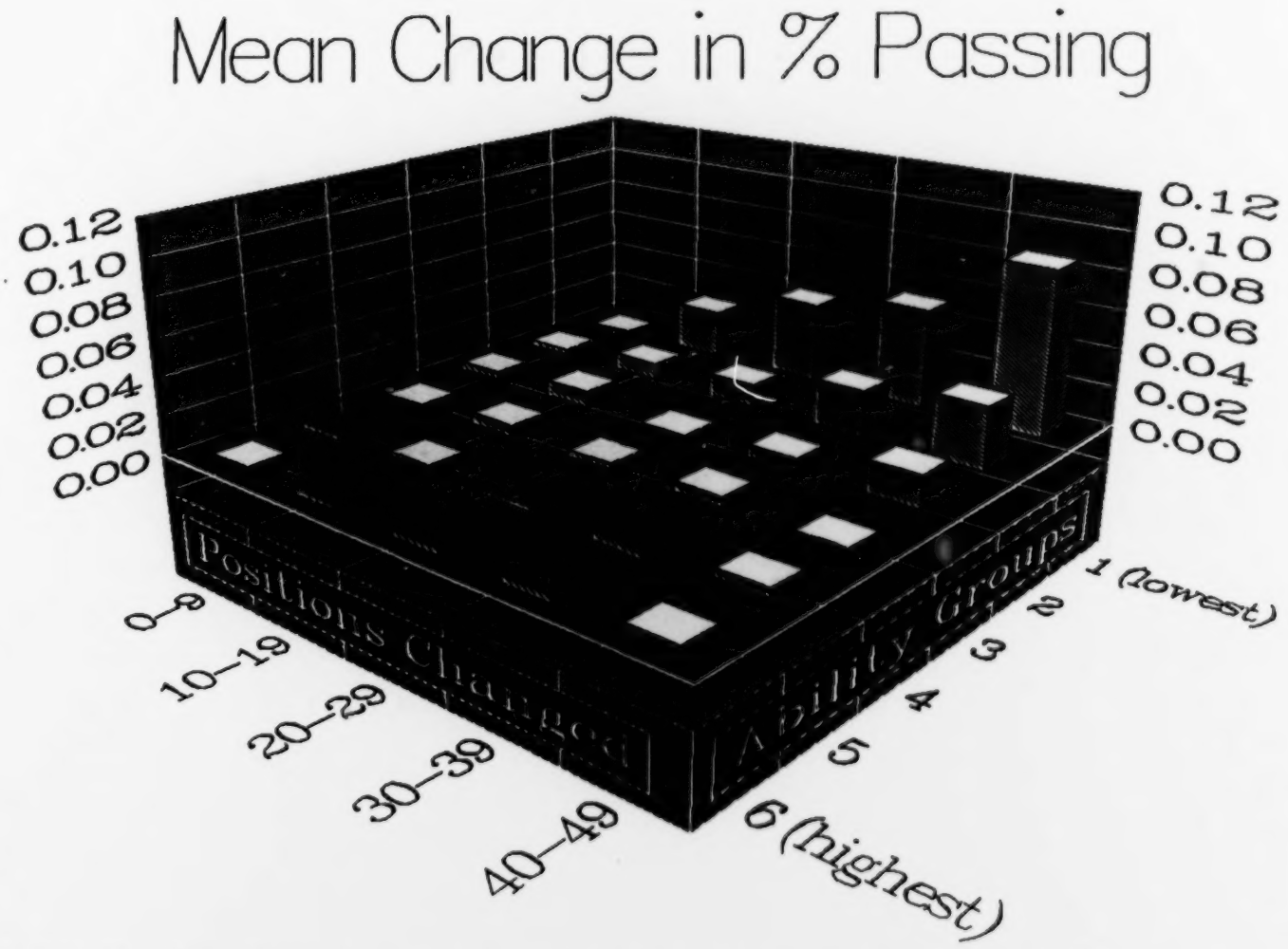

Figure 11. Mean Change in percent passing by ability groups and number of positions changed